Meta-Learning with Selective Data Augmentation for Medical Entity Recognition

Asma Ben Abacha & Dina Demner-Fushman

National Library of Medicine
National Institutes of Health
Bethesda, MD, USA
asma.benabacha@nih.gov; ddemner@mail.nih.gov

Abstract. With the increasing number of annotated corpora for supervised Named Entity Recognition, it becomes interesting to study the combination and augmentation of these corpora for the same annotation task. In this paper, we particularly study the combination of heterogeneous corpora for Medical Entity Recognition by using a meta-learning classifier that combines the results of individual Conditional Random Fields (CRFs) models trained on different corpora. We propose selective data augmentation approaches and compare them with several meta-learning algorithms and baselines. We evaluate our approach using four sub-classifiers trained on four heterogeneous corpora. We show that despite the high disagreements between the individual models on the four test corpora, our selective data augmentation approach improves performance on all test corpora and outperforms the combination of all training corpora.

Keywords: Medical Entity Recognition, Data Augmentation, Meta-Learning

1 Background

Several Natural Language Processing applications such as Question Answering require extracting Named Entities (NE) to obtain reference semantic anchors for the interpretation of natural language [10, 9]. Named Entity Recognition (NER) aims to identify phrases that refer to pre-defined concepts (e.g. Person, Organization, Location) by tagging sequences of words that correspond to NE. Supervised machine learning (ML) methods proved to perform better in NER than rule-based methods or unsupervised methods if a sufficient amount of relevant training data is available [11, 17].

More and more corpora are available for supervised NER in biomedical texts [18, 2]. Corpora with different characteristics have been shown to decrease systems' performance in related work. For instance, for two corpora with the approximately equal numbers of annotations, the F-measure for a Support Vector Machines (SVM) classifier was 18% lower for the corpus with less unique annotations and shorter sentences [15]. Combining AIMed and GENIA corpora

for biomedical entity recognition reduced F-measure by 9.88% and combining GENIA and GENETAG led also to a drop of 3.34% in F-measure [3].

This observation also applies to the medical domain where the same task is often tackled from different perspectives. One of the most salient examples are Medical Entity Recognition (MER) corpora that can consist of clinical texts, scientific abstracts, discharge summaries or encyclopedia articles. Combining these corpora is a challenge because of their linguistic heterogeneity (e.g. use of acronyms vs. full names, specialist vs. public vocabulary or short vs. long sentences) and their annotation bias that could be the result of different guidelines for the definition of entity boundaries (e.g. exclusion/inclusion of modifiers), for the exclusion/inclusion of unspecific entities such as 'treatment', 'injection', 'drug', etc. or nested entities which may or may not be taken into account.

These different characteristics can also cause a significant decrease in performance when combining different training corpora. On the other hand, a successful corpus/model augmentation approach could lead to a significant impact as it will reduce the need to add or expand annotated corpora while enhancing performance in the same time.

Meta-learning proved to be effective in enhancing performance through combinations of different systems and feature sets for the same corpus [16, 3]. However, as far as we know, no successful approach has been proposed to combine models obtained from heterogeneous corpora for MER. In this paper we investigate such combination with two selective data augmentation approaches based on inter-corpus agreements. We conduct experiments on two clinical texts corpora: i2b2'10 [18] and SemEval'15 [12] and two scientific abstracts corpora: NCBI [2] and Berkeley [13]. Statistics on each corpus are presented in Table 1. We compare our selective data augmentation approaches with Conditional Random Fields and different meta-learning algorithms applied to the union of all training corpora. The obtained results show that our data augmentation approach outperforms all individual classifiers and meta-learning baselines.

We present our methods and baselines in the following section. In section 4 we discuss the obtained results which are presented in section 3.

2 Methods

We consider the combination of different models built from heterogeneous training corpora using a meta-learning classifier. We propose new data augmentation methods and compare them with the individual classifiers and two different meta-learning baselines which are presented in this section.

2.1 Individual Classifiers

We represent the NER problem through the IOB labels at token level (i.e. I: Inside, O: Out, B: Beginning). Conditional Random Fields (CRFs) [7] have shown success in many sequence labelling tasks such as parts-of-speech (POS) tagging [14, 6] and NER [8, 5]. We build an individual CRF classifier on each corpus using

Corpus	Genre	Entity Types	Size: training & test corpora	Training: Token Nbr	Training: Pos. Class Balance
i2b2 2010	Discharge summaries and progress notes.	Problem, Treatment, Test	349 files for training + 477 files for test (76,665 sentences).	260,538	10.93
SemEval 2015	Clinical notes from MIMIC II database.	Problem	298 files for training + 133 files for test.	253,066	7.51
Berkeley 2004	Scientific abstracts and titles from MEDLINE.	Problem, Treatment	2,655 sentences for training + 1,000 sentences for test.	47,293	4.22
NCBI 2014	PubMed abstracts.	Problem	693 abstracts for training + 100 abstracts for test.	153,425	7.62

Table 1. Statistics on each corpus.

the same features for each corpus. We note by M_i each individual model built from a training corpus i. The feature set includes:

- Word features: The word itself, 2 preceding words, 3 following words, and their lemmas. An additional feature that indicates whether the current word is a stop word or not.
- Morpho-syntactic features: POS tags of the word itself, 2 previous words and 3 following words. We used Stanford Parser for POS tagging and tokenization [1]. Additional features include: next verb, next noun and next adjective or adverb.
- Orthographic features: Presence of hyphen, plus sign, ampersand or slash. The word is a number, a letter, a punctuation sign or a symbol. The word is in uppercase, capitalized, in lowercase (AA, Aa, aa). Prefixes of different lengths (from 1 to 4). Suffixes of different lengths (from 1 to 4).

2.2 Meta-Learning Approach

We used the results of the individual models as features to build a meta-learning classifier. Individual models are trained and tested on the training corpora only. For meta-learning, we consider the set of features MF:

- the POS tag of the token K.
- the N labels (I, O, or B) predicted by each individual model for K.
- the POS tag of the previous token K_{-1}.
- the N labels (I, O, or B) predicted by each individual model for K_{-1}.

We do not use other word-level features such as the words and lemmas in order to bypass domain adaptation and corpus bias.

We generate a meta-learning dataset d_i from each individual training corpus c_i using the set of features MF. Meta-test datasets are constructed from the original test corpora with the same set of features MF. We train several meta-algorithms on each meta-learning dataset and evaluate them on all meta-test datasets individually.

2.3 Selective Data Augmentation

The goal of selective data augmentation is to select the most relevant data to augment the target meta-learning dataset d_i from other meta-learning datasets $d_j, j \neq i$. By augmenting these datasets we are driving the learning algorithm to take into account the decisions of the individual model(s) that are selected by the augmentation approach.

We consider the NER task as a token classification problem. In that perspective, the intuition of our selective data augmentation approaches is that negative examples (i.e. 'O' class in the IOB format) are best detected with the train portion of the same corpus as it has significantly more examples than the positive classes (i.e. I and B in the IOB format). The goal is therefore to increase the number of positive examples with relevant data from the external heterogeneous corpora in order to reach a positive-class ratio , noted $O^{'}$, higher than the initial ratio, noted O.

We address this problem using the level of agreement between the positive-class labels predicted by the individual models for a given test corpus t. We use the F-measure as the pairwise agreement $A_t(M_1, M_2)$ between the predictions of the individual models M_1 and M_2 for a given test dataset t. In order to have an application-driven (semantic) agreement we compute $A_t(M_1, M_2)$ over entity spans rather than token labels. Table 2 presents agreement values for the four corpora considered in our experiments. To avoid corpus bias, we compute the agreement between two individual models as their average agreement over all test corpora:

$$A(M_1, M_2) = \frac{\sum_{i=1}^{N} A_{t_i}(M_1, M_2)}{N} \quad (1)$$

Where N is the total number of test corpora.

For a given meta-learning dataset d_1, a heterogeneous dataset d_2 is likely to be a relevant source for data augmentation if it has a high agreement value. Our first augmentation approach, called **Most Agreeing auGmEntation ($MAGE$)** relies on this assumption by adding positive examples only from the most agreeing model. The positive examples to be added to a training dataset d_1 are taken randomly from the dataset d_i that satisfies $A(M_1, M_2) = \max_k A_(M_1, M_k)$ until the target positive class ratio $O^{'}_{d_1}$ is reached or until positive examples from d_i are exhausted.

In a second augmentation approach, called **HIgher Ratio augmEntation ($HIRE$)** we use the positive-class ratio of each training dataset as an indicator

Agreement (F-measure %)		**i2b2-test**		**SemEval-Test**		**Berkeley-Test**		**NCBI-Test**	
Model 1	Model 2	**ES**	**IS**	**ES**	**IS**	**ES**	**IS**	**ES**	**IS**
i2b2-train	SemEval-train	**50.90**	**80.14**	**50.61**	**79.58**	41.09	**77.38**	25.36	**60.40**
	Berkeley-train	20.86	28.04	18.41	22.76	**42.68**	63.37	14.49	26.39
	NCBI-train	21.93	35.79	18.50	30.71	34.95	62.28	20.38	51.54
SemEval-train	Berkeley-train	24.40	29.87	19.88	22.92	48.63	68.25	**26.41**	35.94
	NCBI-train	26.97	37.86	21.68	31.19	43.42	67.34	**39.54**	**60.09**
Berkeley-train	NCBI-train	**37.62**	**50.74**	**29.85**	**41.40**	**50.11**	**73.97**	23.54	36.55

Table 2. Agreement (F-measure) between individual models pairwise. Highlighted: first and second best agreement on each test corpus. ES: Exact Span. IS: Inexact Span.

of reliability and restrict data augmentation to the training datasets d_i that satisfy: $O_{d_i} > O_{d_1}$ and $A(M_1, M_i) > D$. D is used as a threshold to avoid using datasets having low agreement with d_1 which may bring more noise than useful data.

From another point of view, datasets with low agreement might be more likely to detect new entities as they carry the highly different perspective of their heterogeneous corpora. In order to benefit from all levels of agreement, we propose a third data selection approach called **Proportional Agreement auGmEntation** ($PAGE$). $PAGE$ selects from each dataset d_i an amount of random data $PAGE_{d_1}(d_i)$ that is proportional to the relative agreement between d_i and the dataset d_1 to be augmented. The relative agreement $RA(d_1, d_i)$ is defined as (N is the number of test corpora):

$$RA(d_1, d_i) = \frac{Exp(A(M_1, M_i))}{\sum_{k=1}^{N} Exp(A(M_1, M_k))} \tag{2}$$

Equation 3 presents the exact formula for $PAGE_{d_1}(d_i)$:

$$PAGE_{d_1}^{O_{c_1}^{'}}(d_i) = R_d \times \frac{Exp(A(M_1, M_i))}{\sum_k Exp(A(M_1, M_k))} \times (P_{d_1}^{'} - P_{d_1}) \tag{3}$$

- $P_{d_1}^{'}$ represents the number of positive examples needed to reach the target positive-class ratio $O_{c_1}^{'}$.
- P_{d_1} is the number of initial positive examples in d_1.
- Exp is used to highlight further the gaps in agreement and avoid drowning the most agreeing corpus by the total number of corpora.
- R_d is a reduction factor used to keep the same initial proportions if the positive examples from a dataset d_i are exhausted before reaching $PAGE_{d_1}^{O_{d_1}^{'}}(d_i)$. It is computed automatically at runtime.

The training set obtained after augmentation is given as input to a Bagging algorithm. Bagging was selected because it provided a slightly better or similar performance when compared to several other algorithms including Neural Networks, Dagging, Stacking, Multi-Boost and Ada-Boost.

3 Experiments

We consider two meta-learning baselines:

- A Conditional Random Fields model: All_{CRF}, built from the union of all training corpora using the same features set as the individual classifiers.
- The best-performing meta-learning algorithm among Bagging, Dagging, Stacking, Multi-Boost and Ada-Boost: All_{meta}, trained on the union of all training corpora with the set of features MF defined is section 2.2.

In this section, we evaluate the ability of the meta-learning baselines and the selective data augmentation approaches in the identification of named entities referring to medical problems. The evaluation measures are: Precision, Recall, and F-measure. We use the four corpora described in Table 1 that have different genres (clinical text vs. scientific abstracts), different annotation rules, different sizes and different positive-class ratios. We first evaluate each individual model on the four test corpora (cf. Table 3) then present the results of the different augmentation approaches in Table 4. We used the Weka platform [4] to compare different algorithms with our augmented datasets approach. In these experiments we set the target class balance $O^{'}$ to 20 and the agreement threshold D to 0.5 for the $HIRE$ approach. We discuss the impact of different values for $O^{'}$ and D in section 4.

F-measure (%)		**i2b2-Test**		**SemEval-Test**		**Berkeley-Test**		**NCBI-Test**	
		ES	**IS**	**ES**	**IS**	**ES**	**IS**	**ES**	**IS**
I2b2-train	**R**	**74.9**	**83.7**	41.9	72.7	**58.0**	**87.1**	24.9	62.0
	P	**84.8**	95.0	47.9	81.0	44.1	67.2	19.9	48.1
	F	**79.59**	**89.03**	44.74	76.66	**50.1**	**75.9**	22.1	54.1
SemEval-train	**R**	46.0	73.2	**64.5**	**72.7**	40.6	75.3	34.8	54.1
	P	55.2	91.8	**75.8**	85.5	38.3	73.8	41.9	65.1
	F	50.2	81.4	**69.7**	**78.6**	39.4	74.5	38.0	59.1
Berkeley-train	**R**	11.3	15.5	09.1	11.4	38.6	57.6	12.49	20.0
	P	70.2	**96.9**	75.5	**94.5**	**58.5**	**86.6**	60.9	**97.9**
	F	19.6	26.7	16.25	20.3	46.5	69.1	20.72	33.3
NCBI-train	**R**	12.5	19.8	11.3	16.2	32.06	53.5	**75.1**	**82.6**
	P	56.5	93.8	57.7	82.6	48.3	82.4	**86.0**	94.6
	F	20.5	32.7	18.9	27.1	38.5	64.8	**80.2**	**88.2**

Table 3. The impact of training and testing on different corpora: Comparison of the performance of individual classifiers. ES: Exact Span. IS: Inexact Span.

4 Discussion

We tested 11 target ratio $O^{'}$ values in $[10, 20]$ and four more values in $\{30, 40, 50, 60\}$. The target positive-class ratio $O^{'}$ of 20 led to best results in all 3 augmentation

F-measure (%)		**i2b2-Test**		**SemEval-Test**		**Berkeley-Test**		**NCBI-Test**	
		ES	**IS**	**ES**	**IS**	**ES**	**IS**	**ES**	**IS**
Ind.	**R**	74.9	83.7	64.5	72.7	38.6	57.6	75.1	82.6
	P	**84.8**	95.0	**75.8**	85.5	**58.5**	86.6	86.0	94.6
	F	79.5	**89.0**	69.7	78.6	46.5	69.1	80.2	88.2
$All_{(CRF)}$	**R**	66.5	82.0	58.5	74.6	53.7	82.7	74.0	83.1
	P	76.9	**95.7**	67.8	**85.9**	50.9	78.2	82.9	93.1
	F	71.3	88.4	62.8	79.8	52.2	**80.4**	78.2	87.8
$All_{(meta)}$	**R**	41.9	73.3	40.3	55.2	53.0	82.2	79.7	88.3
	P	49.8	87.0	46.8	64.1	38.1	59.2	54.2	60.1
	F	45.5	79.6	43.3	59.3	44.3	68.9	64.5	71.5
$MAGE^{20}$	**R**	75.3	**89.0**	**77.4**	**84.2**	47.6	80.4	77.5	88.1
	P	73.8	87.3	71.1	77.3	56.3	68.0	62.9	71.6
	F	74.5	88.1	**74.1**	**80.6**	51.6	73.7	69.4	79.0
$PAGE^{20}$	**R**	71.22	**89.0**	77.3	**84.2**	55.9	82.9	76.2	88.4
	P	70.5	88.2	69.6	75.8	46.1	68.4	63.3	73.5
	F	70.8	88.6	73.3	79.8	50.5	75.0	69.2	80.3
$HIRE^{20}_{0.5}$	**R**	76.6	84.0	**77.4**	**84.2**	**68.3**	**90.5**	**78.5**	**83.6**
	P	84.7	92.8	71.1	77.3	45.0	59.7	**88.8**	**94.6**
	F	**80.4**	88.2	**74.1**	**80.6**	**54.3**	72.0	**83.3**	**88.8**

Table 4. Corpus Augmentation Results. Parameters: $O^{'} = 20$, $D = 0.5$.

methods. With that setting, all positive examples from the selected corpora were exhausted for the 3 methods (values higher than 20 led to the exact same results). Lower values of $O^{'}$ (< 20) led to much lower performance depending on the method and corpus. This can be explained by the random data subset selection which may not lead to sufficiently correlating feature-vector patterns. It can also be explained by the fact that agreement levels were computed over the full models that used all of their respective examples.

HIRE provided the best performance compared to both the individual models and the other data augmentation methods. This is supported by our observations from the output of both *MAGE* and *PAGE*, where using corpora with lower class ratio proved to drop performance in all cases. Table 5 shows the corpora selected for augmentation in each method. The best results for *HIRE* were obtained by our first attempt to set an agreement threshold of 50%. Starting from a threshold lower than 49% performance decreases significantly for NCBI as it is then augmented from the SemEval dataset (agreement of 49.1%). This threshold needs to be tested with more corpora and different entity types to validate further its empiric relevance.

PAGE was the worst performing augmentation method. This particularly shows the sensitivity of the meta datasets to noisy data. For instance the performance on NCBI dropped by 15 points by adding 9% more instances extracted randomly from the SemEval dataset. The same observation can be made on the

augmentation of the i2b2 dataset, where performance dropped significantly with only 1.0% more examples extracted from the Berkeley dataset (cf. Table 5).

Source Corpus	Corpora Selected for Augmentation		
	$MAGE^{20}$	$PAGE^{20}$	$HIRE^{20}_{0.5}$
i2b2	SemEval	SemEval(67.2%), Berkeley (1.0%), NCBI(1.0%)	$\emptyset$
SemEval	i2b2	i2b2(148.6%), NCBI (18.3%), Berkeley(8.8%)	i2b2
Berkeley	NCBI	NCBI(341.5%), SemEval(97.3%), i2b2(26.04%)	NCBI
NCBI	Berkeley	Berkeley(17.2%), SemEval(9.9%), i2b2(1.2%)	$\emptyset$

Table 5. Corpora selected to augment each dataset according to augmentation method. The percentage is computed w.r.t. the number of positive examples in the source dataset. Lack of percentage means all positive examples from the corpus have been used for augmentation.

Overall, our experiments show that:

- F-measure based agreement computed over a set of several heterogeneous test corpora is effective in ordering the datasets according to their relevance for data augmentation,
- a random subset selection is ineffective if the agreement level is low (even for very small subsets),
- corpora with (significantly) lower class balance are not good candidates for augmentation.

From additional experiments, we found that using the agreement ratios as features for the meta-algorithm did not have any impact on the results, as well as several nonlinear combinations of these ratios. We attribute this lack of influence mainly to the lack of variability in the set of values of the agreement measures (or their combination) and also in the recurring feature-vector patterns at token level.

5 Conclusion

Our approach allows benefiting from heterogeneous corpora for Named Entity Recognition. Our experiments on four different corpora showed that selective training data augmentation works better than the combination of all corpora. F-measure was improved on all corpora with an average increase of 4.3%. We also found that the global agreement level between label predictions of the pairwise models is an effective metric in selecting relevant sources for data augmentation when used with potential reliability indicators such as the class balance of each corpus. In order to gain further insights, we also plan to conduct these experiments with other entity types and corpora. While word features did not allow combining the corpora (cf. CRF results in Table 4) the semantic abstractions of each word might be an approach to building a suitable feature set for data augmentation in the future.

Acknowledgements

This research was supported by the Intramural Research Program at the U.S. National Library of Medicine, National Institutes of Health.

References

1. Chen, D., Manning, C.: A fast and accurate dependency parser using neural networks. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 740–750. Association for Computational Linguistics, Doha, Qatar (October 2014), http://www.aclweb.org/anthology/D14-1082
2. Dogan, R.I., Leaman, R., Lu, Z.: NCBI disease corpus: A resource for disease name recognition and concept normalization. Journal of Biomedical Informatics 47, 1–10 (2014), http://dx.doi.org/10.1016/j.jbi.2013.12.006
3. Ekbal, A., Saha, S., Sikdar, U.K.: Biomedical named entity extraction: some issues of corpus compatibilities. SpringerPlus 2 (2013)
4. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H.: The weka data mining software: An update. SIGKDD Explor. Newsl. 11(1), 10–18 (Nov 2009), http://doi.acm.org/10.1145/1656274.1656278
5. He, Y., Kayaalp, M.: Biological entity recognition with conditional random fields. In: AMIA 2008, American Medical Informatics Association Annual Symposium, Washington, DC, USA, November 8-12, 2008 (2008), http://knowledge.amia.org/amia-55142-a2008a-1.625176/t-001-1.626020/f-001-1.626021/a-060-1.626384/a-061-1.626381
6. Huang, Z., Xu, W., Yu, K.: Bidirectional LSTM-CRF Models for Sequence Tagging. ArXiv e-prints (Aug 2015)
7. Lafferty, J.D., McCallum, A., Pereira, F.C.N.: Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In: Proceedings of the Eighteenth International Conference on Machine Learning (ICML 2001), Williams College, Williamstown, MA, USA, June 28 - July 1, 2001. pp. 282–289 (2001)
8. Lee, C., Hwang, Y.G., Oh, H.J., Lim, S., Heo, J., Lee, C.H., Kim, H.J., Wang, J.H., Jang, M.G.: Fine-grained named entity recognition using conditional random fields for question answering. In: Ng, H., Leong, M.K., Kan, M.Y., Ji, D. (eds.) Information Retrieval Technology, Lecture Notes in Computer Science, vol. 4182, pp. 581–587. Springer Berlin Heidelberg (2006), http://dx.doi.org/10.1007/11880592_49
9. Mendes, A.C., Coheur, L., Lobo, P.V.: Named entity recognition in questions: Towards a golden collection. In: Chair), N.C.C., Choukri, K., Maegaard, B., Mariani, J., Odijk, J., Piperidis, S., Rosner, M., Tapias, D. (eds.) Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10). European Language Resources Association (ELRA), Valletta, Malta (may 2010)
10. Mollá, D., Zaanen, M.V., Smith, D.: Named entity recognition for question answering. In: In Lawrence Cavedon and Ingrid Zukerman, editors, Proceedings of the 2006 Australasian Language Technology Workshop. pp. 51–58 (2006)
11. Nadeau, D., Sekine, S.: A survey of named entity recognition and classification. Linguisticae Investigationes 30(1), 3–26 (2007)

12. Pradhan, S., Elhadad, N., South, B., Martinez, D., Christensen, L., Vogel, A., Suominen, H., Chapman, W., Savova, G.: Evaluating the state of the art in disorder recognition and normalization of clinical narrative. Journal of the American Medical Informatics Association (JAMIA) 22(1), 143–154 (2015)
13. Rosario, B., Hearst, M.A.: Classifying semantic relations in bioscience texts. In: Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics, 21-26 July, 2004, Barcelona, Spain. pp. 430–437 (2004), http://acl.ldc.upenn.edu/acl2004/main/pdf/309_pdf_2-col.pdf
14. Sha, F., Pereira, F.C.N.: Shallow parsing with conditional random fields. In: HLT-NAACL (2003), http://acl.ldc.upenn.edu/N/N03/N03-1028.pdf
15. Shmanina, T., Zukerman, I., Yepes, A.J., Cavedon, L., Verspoor, K.: Impact of corpus diversity and complexity on ner performance. In: Australasian Language Technology Association Workshop 2013. p. 91 (2013)
16. Si, L., Kanungo, T., Huang, X.: Boosting performance of bio-entity recognition by combining results from multiple systems. In: Proceedings of the 5th International Workshop on Bioinformatics. pp. 76–83. BIOKDD '05, ACM, New York, NY, USA (2005), http://doi.acm.org/10.1145/1134030.1134044
17. Tkachenko, M., Simanovsky, A.: Named entity recognition: Exploring features. In: 11th Conference on Natural Language Processing, KONVENS 2012, Empirical Methods in Natural Language Processing, Vienna, Austria, September 19-21, 2012. pp. 118–127 (2012), http://www.oegai.at/konvens2012/proceedings/17_tkachenko12o/
18. Uzuner, Ö., South, B.R., Shen, S., DuVall, S.L.: 2010 i2b2/va challenge on concepts, assertions, and relations in clinical text. JAMIA 18(5), 552–556 (2011), http://dx.doi.org/10.1136/amiajnl-2011-000203